Anchor Text의 단어 정보를 이용한 자동 문서 범주화

허희근[1], 한기덕[2], 정성원[3], 임성신[4], 권혁철[5]
*부산대학교 정보컴퓨터공학부
{riniya[1], templer[2], swjung[3], sslim[4], hckwon[5]} @pusan.ac.kr

Automatic Text Categorization Using Term Information of Anchor Text

Hee-keun Heo[1], Gi-deok Han[2], Sung-won Jung[3],
Sung-shin Lim[4], Hyuk-chul Kwon[5]
*Dept. of Computer Science and Engineering,
Pusan National University

요 약

최근의 웹 문서는 텍스트뿐만 아니라 이미지, 사운드 등 다른 여러 형태로 표현되고 있어서 텍스트의 비중이 낮아지고 있다. 그래서 문서 내에서 일정량 이상의 단어 추출이 어려운 문서들에 대해서 기존의 단어 정보만을 이용한 문서 범주화 방법은 좋은 성능을 기대할 수 없다. 그래서 본 논문은 Anchor Text 단어 정보의 자질 적합성 판단에 의한 새로운 자동 문서 범주화 모델을 제안한다. 문서 범주화 모델로는 베이지언 확률 모델을 이용하였으며, 카이제곱 통계량을 사용하여 자질을 선정하였다. 문서 내에서 추출된 단어 자질들이 해당 문서를 판단하는데 부족하다고 판단되면 문서의 링크정보를 이용하여 연결된 문서의 단어 자질과 Anchor Text의 단어 자질을 반영함으로써 성능을 향상시킨다.

1. 서론

문서 범주화에 대한 연구는 효과적인 범주화 방법과 문서 자질의 추출방법을 중심으로 발전하였다. 범주화 방법으로는 기계학습 모델, 확률 모델, 정보검색 모델 등을 적용하여 다양한 연구가 진행되고 있으며[7], 범주 대표어로 사용되는 자질의 추출방법은 학습 문서 내의 명사들을 추출하고, 자질의 출현 빈도를 이용하여 해당 자질의 가중치를 계산하는 것이 일반적인 방법이다.

최근의 웹 문서는 텍스트뿐만 아니라 이미지, 사운드 등 다른 여러 형태로 표현되고 있어서 텍스트의 비중이 낮아지고 있다. 그래서 문서 내에서 일정량 이상의 단어 추출이 어려운 문서들에 대해서 기존의 단어 정보만을 이용한 문서 범주화 방법은 좋은 성능을 기대할 수 없다. 이를 해결하기 위해서 웹 문서의 링크정보를 이용하여 정보검색의 성능을 향상시키려는 연구가 활발히 진행되고 있다.

본 논문에서는 이러한 연구들을 바탕으로 새로운 문서 범주화 모델을 제안한다. 이 모델은 웹 문서 내에서 추출한 단어 자질이 해당문서를 판단하는데 부족하다고 판단되면, 문서의 링크정보를 이용하여 연결된 문서의 단어 자질과 Anchor Text의 단어 자질을 반영함으로써 성능 향상을 기한다. 특히, 링크 정보를 이용하여 연결된 문서의 단어 자질을 반영하기 위해 문서를 확장할 때, 같은 도메인에 있는 문서들을 반영했다. 그 이유는 다른 도메인에 링크된 문서에 비해 단어의 연관성과 유사성이 높기 때문이다.[6] 또한 본 논문이 제안하는 모델과 기존의 카이제곱 통계량과 나이브 베이즈 분류기만을 이용한 범주화 모델을 비교하였다.[8]

본 논문의 구성은 다음과 같다. 2장에서 문서 범주화에 대한 관련연구를 살펴보고, 3장에서는 문서범주화 모델의 구조와 자질을 선별하는 방법과 학습 모델을 설명한다. 4장에서는 실험 및 결과를 평가하고, 5장에서 결론 및 향후 과제에 대해서 기술한다.

2. 관련 연구

문서 범주화는 사전 분류체계에 기초하여 분류대상이 되는 문서를 가장 적합한 범주에 할당함으로써 문서들의 집단을 형성하는 기법으로, 범주를 대표할 수 있는 단어 자질을 선정하고 각 자질들에 대한 가중치를 부여하여 문서를 미리 정의되어 있는 범주로 할당시키는 방법이다[8].

유사한 웹 문서가 포함된 범주를 대표하는 단어 자질을 선택하거나 단어 자질의 가중치를 계산하기 위한 기법이 있는데, 단어의 단순 출현 빈도에 의해 가중치를 계산하는 방법을 보완하기 위해 단어의 출현 빈도(TF : Term Frequency)를 기반으로 역문헌

빈도(IDF : Inverse Document Frequency), 역범주 빈도(ICF : Inverse Collection Frequency)[12]를 사용한다. 단어 자질의 선택은 웹 문서에 나타나는 단어들 중에서 범주를 구분하는데 사용될만한 단어들을 선택하는 작업이다. 학습문서에서 나타나는 단어의 수가 많기 때문에 모든 단어가 자질로 선택된다면 학습 및 분류 시간이 매우 오래 걸리게 된다. 따라서 문서 범주화의 성능의 저하 없이 자질의 수를 줄이기 위해 문서에 나타나는 단어의 정보량을 계산하고, 정보량이 큰 단어만을 자질로 선택하려는 연구가 활발히 진행되어 왔다. 범주를 대표할 수 있는 중요한 자질을 선별하는 기법으로 카이제곱 통계량, 상호 정보량(Mutual Information)[2], 기대 상호 정보량(Expected MI) 정보 획득량(Information Gain)[11] 등이 있다.

문서 범주화 모델은 베이지언 확률모델[3]과 결정트리 모델[9], 지지 벡터 기계, k-최근린법, 선형분리기 등이 있다.

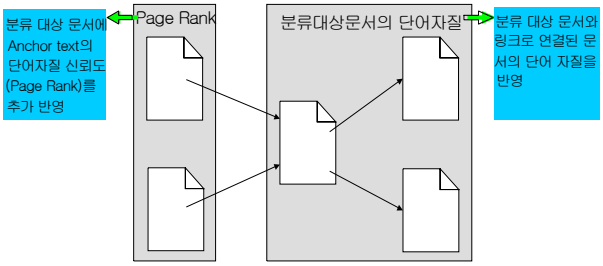
3. Anchor Text의 단어 정보를 이용한 자동 문서 범주화

3-1 색인어 벡터 만들기

색인어 벡터는 다음 과정을 통해 만든다.

부산대학교 한국어 정보처리 연구실의 웹로봇을 사용하여 1000만 건의 문서를 수집하였다. 그리고 색인과정을 거친 후, Anchor Text의 색인어 정보를 벡터로 만든다. 색인어 정보 벡터는 Page Rank, Anchor Text, 연결된 문서 간의 링크 정보를 사용하여 만든다.

[그림 1]은 본 논문이 제안하는 범주화 모델이다.



[그림 1] 범주화 모델

3-2 문서 자질의 적합성 판단 임계치 설정

색인과정을 통해 추출된 단어를 사용하여 해당 문서의 벡터를 구성한다. 이때 해당 문서 내부의 출현 단어 수가 문서 범주화에 적합한지의 여부를 결정해야 한다. 이러한 판단을 위해서는 문서 범주화에 적합한 문서에 출현하는 단어 수의 임계치가 필요하며, 이를 위해 문서에 출현 단어 수와 문서 범주화의 모델의 성능관계를 조사한 후 적절한 임계치를 설정할 수 있도록 다음과 같은 실험을 하였다. 실험대상은 검색 포탈 사이트인 네이버 디렉토리(http://dir.naver.com/)의 "자연과학" 범주 내의 하위 6개 범주의 문서를 선정하였다. 그 이유는 적절한 분류기준에 따라 전문가에 의해 수작업으로 분류된 문서와 실험에 의해 분류된 문서를 비교함으로써 문서분류 성능에 따른 임계치를 쉽게 구할 수 있기 때문이다. 또한, '자연과학' 분야의 경우 웹 문서가 풍부하고 범주 내의 하위 범주 구분이 명확하기 때문에 실험 대상으로 적합하다고 판단하였다.

분석대상으로 선정된 범주와 대상 문서 수는 [표 1]과 같다.

범 주	문서 수
농학	145
물리학	102
생물학	226
수학	102
천문학	223
화학	113
계	911

[표 1] 범주별 분석 대상 문서

실험에 사용되는 학습문서는 단어 자질이 충분한 문서를 위주로 1200건을 수집하였다. 분류자질의 선별단계에서는 카이제곱 통계량, 문서 분류단계에서는 베이즈언 확률 모델을 사용하여 문서를 분류하였다.

실험을 통해서 문서 내에 출현하는 단어의 수가 60개 이상일 경우 문서 범주화 모델이 추출된 단어 수에 영향을 받지 않는다는 것을 확인할 수 있었다. 반면에 60개 미만인 문서는 전체 범주화 모델의 성능 저하에 결정적인 영향을 주었고, 이런 문서의 경우는 링크정보를 이용하여 연결된 문서 중에서 단어의 수가 60개 이상인 문서를 반영하고, 반영된 문서의 수로 원래 문서의 가중치를 정규화시킨다.

3-3 카이제곱 통계량을 적용한 자질 선별

자질선별이란 범주를 대표할 만한 중요한 용어를 얻는 방법이다. 일반적으로 카이제곱 통계량과 정보획득량이 좋은 성능을 보인다[2]. 특히, 카이제곱 통계량은 중요 자질을 순위화하여 벡터 차원을 줄일 수 있으며, 용어 t와 범주 c와의 의존성을 측정하는데 사용된다. 계산하는 식은 식 (1)과 같다.[8]

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

A : 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있는 문서 수

B : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서 수

c : 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있지 않은 문서 수

d : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있지 않은 문서 수

N : 학습에 사용된 전체 문서 수

카이제곱 통계량은 용어 t와 범주 c가 완전히 독립적이면 0의 값을 가진다. χ^2값으로 선별된 자질들을 범주별 하나의 자질로 선별하는 작업이 필요하며 식 (2)와 같다.[8]

$$\chi^2 \max(t) = \max\ \chi^2(t, c_1), \chi^2(t, c_2), \ldots, \chi^2(t, c_n) \quad (2)$$

식 (2)을 이용하여 자질에 대한 유일한 값을 중요도별로 순위화할 수 있다.

3-4 베이지언 분류 모델

본 논문에서는 단순 베이지언(Naive Bayesian) 모델을 이용하여 문서를 분류하였다. 이 모델은 대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률 값을 갖는 범주에 그 문서를 할당하는 방법이다. 계산하는 식은 식(3)과 같다.

$$Max_c[P(c \mid d) = Max_c\left[\frac{P(c)(d \mid c)}{P(d)}\right] = Max_c\left[P(c)\prod_{i=1}^{T}P(t_i \mid c)^{N(t_i \mid d)}\right] \quad (3)$$

$$P(c)\prod_{i=1}^{T}P(t_i \mid c)^{N(t_i \mid d)} \propto \frac{\log P(c)}{n} + \sum_{i=0}^{T}P(t_i \mid d)\log\left(\frac{P(t_i \mid c)}{P(t_i \mid d)}\right) \quad (4)$$

식 (3)의 $N(t_i \mid d)$는 문서 d에서 용어 t_i가 출현하는 횟수(TF)를 의미하고 T는 전체 문서 집합 내의 용어 수를 나타낸다. 그러나, $N(t_i \mid d)$가 커질수록 오히려 $P(c \mid d)$값이 작아지므로 이 문제를 해결하기 위해 식 (4)로 변형했다. 식 (4)의 $P(c)$는 전체 학습문서 집합에서 해당 범주가 나타날 확률을 의미하고, $P(t_i \mid c)$는 해당 범주에서 t_i가 출현할 확률, $P(t_i \mid d)$는 대상 문서에서 t_i가 출현할 확률을 의미한다. 문서 간의 차이를 나타내기 위해 Kulback-Leiber Divergence 값을 사용하였고, 각각의 범주에 대한 KL Divergence 값을 표현하기 위해 $P(t_i \mid c)$를 $P(t_i \mid d)$로 나누었다.[14] 각각의 확률 값은 다음과 같이 구할 수 있다.[15]

$$c = vec(w_1, w_2, \ldots, w_T)$$

$$w_1 = P(t_i \mid c) = \begin{cases} \text{최소확률값} & iff(t_i \mid c) = 0 \\ \dfrac{N(t_i \mid c) + 0.5}{Total\_N(c) + 0.5 \times T(c)} & iff(t_i \mid c) \neq 0 \end{cases} \quad (5)$$

$Total\_N(c)$: 범주 c에 나타난 전체 용어 출현 빈도
$T(c)$: 범주 c에 나타난 용어의 수

범주 c는의 w_1벡터로 표현되는데, w_1은 범주 c에서 i번째 용어 t_i의 가중치를 의미하는 것으로 식 (4)의 $P(t_i \mid c)$에 해당한다. 최소 확률값은 0.0000001이다. 마찬가지로 문서 d는 용어 t_i의 가중치 w_1을 갖는 벡터로 표현되며 식 (4)의 $P(t_i \mid d)$에 해당한다.

$$d = vec(w_1, w_2, \ldots, w_T)$$

$$w_1 = P(t_i \mid d) = \frac{N(t_i \mid d) + 0.5}{Total\_N(d) + T(d)} \quad (6)$$

4. 실험 및 결과

4-1 실험데이터 선정

실험대상은 검색 포탈 사이트인 네이버 디렉토리(http:// dir.naver.com/)의 "학문,과학" 범주 내의 하위 5개 범주의 문서를 선정하였다. 실험에 사용할 전체 범주는 [표 2]와 같고, 학습 문서와 테스트 문서는 임의로 선택하였다.

범 주	학습 문서	테스트 문서	총 계
건축공학	200	100	300
기계공학	150	75	225
법 학	200	100	300
생 물 학	200	100	300
전기전자공학	180	90	270
총 계	930	465	1395

[표 2] 실험에 사용된 범주표

[표 2]의 문서는 "학문,과학" 범주 내의 모든 문서 중에서 부산대 한국어 정보처리 연구실의 웹 로봇으로 수집한 1000만 건의 문서 내에 포함된 문서들만 이용하였다. 모든 학습문서와 테스트 문서는 문서에서 추출된 단어의 수가 60개 미만이라면 대상 문서와 링크로 연결된 사이트 내부 문서에서 단어를 추출하여 대상문서의 가중치를 조정하였다.

4-2 실험결과

아래의 [표 3]은 카이제곱 통계량을 이용해 선별된 각 범주별 자질의 예이다.

범 주	자질의 예
건축공학	건축, 건축학, 설계, 건축과, Architecture, 건축사, 인테리어, 건축공, 건축공학, ...
기계공학	기계, 기계공, 설계, 역학, 유체, Mechanical, 기계공학, 동역학, ...
법 학	법, 법률, 법학, 법학과, 형사, 소송, 헌법, 사법, 민사, 법령, ...
생 물 학	생물, 식물, 동물, 생물학, 곤충, 세포, 유전, 생명, 생태, 조류, 미생물 .....
전기전자공학	전자, 전기, 전자공학, 반도체, 전기공학, 소자, 전력, 집적, 전압, 마이크로, 콘덴서, 회로, ...

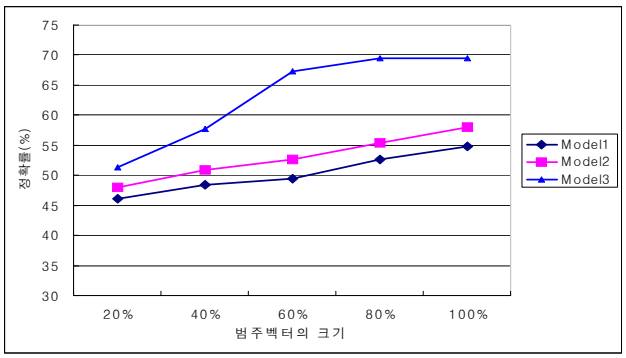
[표 3] 범주별 선택 자질의 예

본 논문에서는 기존의 카이제곱 통계량과 나이브 베이즈 분류기만을 이용한 기존의 범주화 모델과 문서 자질 적합성 판단에 의한 분류 모델을 범주 벡터의 크기를 달리하여 실험하였다.

아래의 [표 4]와 [그림 2]는 범주화 모델에 자질적합성의 판단과 이에 따른 자질을 추가반영한 모델과 그렇지 않은 모델에 대해 범주벡터의 크기를 달리하였을 때의 성능을 비교한 실험의 결과이다. 여기에서 범주할당 성공률이란 전체 문서 중에서 정확히 범주가 할당된 문서들이 차지하는 비율을 나타낸 것이다. 그리고 [그림 2]의 범주벡터 크기의 비율은 범주에 반영되는 문서들의 총 자질의 수를 비율로 나타낸 것이다.

자질비율 \ Model	Model1	Model2	Model3
20%	46.072	47.983	51.38
40%	48.408	50.955	57.749
60%	49.469	52.654	67.304
80%	52.654	55.414	69.427
100%	54.777	57.962	69.427

[표 4] 범주화 모델과 범주벡터의 크기와의 성능 관계



[그림 2] 범주벡터 크기에 따른 모델별 성능

[표 4]의 Model1은 단순히 카이제곱 통계량과 나이브 베이즈 분류기를 이용한 문서 분류 모델이다. Model2와 Model3은 Model1에 자질 적합성 판단을 통한 자질을 추가 반영한 모델로서 Model2는 문서의 링크정보를 이용하여 연결된 문서의 단어 자질을 반영한 모델이고, Model3은 Model2에 Anchor Text의 단어 자질을 추가하여 반영한 모델이다.

실험을 통해 범주화 모델에 자질 적합성의 판단과 이에 따른 자질을 추가반영한 모델이 그렇지 않은 모델보다는 성능이 좋다는 것을 알 수 있었다.

본 논문이 제시한 Model3은 Model1보다 12.78%, Model2보다 10.06%의 성능 향상을 보였다. 범주벡터의 크기가 커지면 커질수록 성능이 향상되지만 사용되는 공간에 비해 성능의 향상도는 그다지 높지 않다. 대량의 문서를 자동으로 분류하기 위해서는 가능한 한 최소의 공간을 이용하는 것이 요구되기 때문에 분류 대상이 되는 문서의 특성을 고려한 처리방법이 바람직하다.

5. 결론 및 향후 과제

자동 문서 범주화(automatic text categorization)는 특히 정보검색 분야와 자연어 처리 시스템에서 그 중요성이 증대되고 있다. 이는 미리 정의된 두 개 이상의 범주에 대해서 입력된 문서의 범주를 자동으로 할당해 주는 작업이다. 수작업으로 문서를 분류하는 한계를 극복하고, 검색성능을 향상시켜 사용자가 원하는 정보를 보다 접근하기 쉽도록 하기 위한 목적으로 현재 다양한 연구와 실험이 계속 진행되고 있다.

본 논문은 Anchor Text 단어 정보의 자질 적합성 판단에 의한 자동 문서 범주화 모델을 제안하였다. 실험을 통하여 카이제곱 통계량과 나이브 베이즈 분류기 모델을 사용하여 자질 적합성 판단에 따른 자질을 추가하는 방법이 효과가 있음을 확인하였다.하지만, 다섯 가지 범주만을 대상으로 실험이 이루어졌다는 점은 이 실험의 한계이다.

향후과제는 분류의 성능을 더 높이기 위해 여러 범주의 자질이 될 수 있는 용어의 범주 모호성을 해소하기 위한 연구도 필요하다.

참고문헌

[1] Yiming Yang, Xin Liu, “A re-examination of text categorization methods”, Proceedings of Conference on Research and Development in Information Retrieval(ACM SIGIR'99), pp.42-99, 1999.

[2] Yiming Yang, Jan O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization", Proceedings of 14th International Conference on Machine Learning, pp.412-420, 1997.

[3] David D. Lewis, “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", Proceedings of 10th European Conference on Machine Learning(ECML'98), 1998.

[4] Yiming Yang, Sean Slattery, Rayid Ghani, “A Study of Approaches to Hypertext Categorization", Journal of Intelligent Information Systems.

[5] Daniele Riboni, “Feature Selection for Web Page Classification", EURASIA-ICT 2002 Proceedings of the Workshops

[6] Brian D. Davison. “Topical locality in the web", Proc. of ACM SIGIR '00, pages 272 - 279, 2000.

[7] 고영중, 서정연, “문서관리를 위한 자동 문서 범주화에 대한 이론 및 기법”, 정보관리 연구, vol. 33, no. 2, pp.19-32, 2002.

[8] 이경찬, 강승식, “자질 중요도 계산 기법에 의한 자동 문서 범주화”, 제15회 한글 및 한국어 정보처리 학술대회, pp.14-17, 2003.

[9] David D. Lewis, Marc Ringuette, “A Comparison of Two Learning Algorithms for Text Categorization", Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[10] M. Steinbach, G. Karypis, and V. Kumar. “A comparison of document clustering techniques". In KDD Workshop on Text Mining, 2000

[11] Tom M. Mitchell, “Machine Learning", McGraw-Hill, 1997.

[12] Salton, G., Fox. E. A., Wu. H., “Extened Boolean Information Retrieval", ncstrl. cornel, pp.82-511, 1982.

[13] Chidanand Apte, Fred Damerau, and Sholom M. Weis, “Towards language independent automated learning of text categorization models,", Proceeding of the 17th annual international ACM-SIGIR, 1994.

[14] L. Douglas Baker and Andrew K. Maccallum, “Distributional Clustering of Words for Text Classification", Proc. of the 21th Annual International ACM-SIGIR, 1998.

[15] Hang Li and Kenji Yamanishi, “Document Classification Using a Finite Mixture Model", The Association for Computational Linguistics, ACL '97, 1997

[16] 이원희, 이교운, 박흠, 김영기, 권혁철, “웹 문서의 단어정보와 링크정보 결합을 이용한 클러스터링 기법”, “제15회 한글 및 한국어 정보처리 학술대회, pp.101-107, 2003.